АНАЛИЗ МЕТОДОВ И АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ПРИМЕНИТЕЛЬНО К ЗАДАЧЕ ВЫБОРА СКВАЖИН-КАНДИДАТОВ НА ПРОВЕДЕНИЕ ГЕОЛОГО-ТЕХНИЧЕСКИХ МЕРОПРИЯТИЙ

Васильева Е.Е.
Томский политехнический университет, Институт кибернетики
katya-vas@mail.ru

Введение

В виду высокой потребности в таком виде сырья как нефть и газ особую актуальность приобретают методы увеличения их добычи. Разработка месторождений требует качественной работы геологов и больших затрат ресурсов. В связи с этим нефтегазодобывающие предприятия пытаются достичь увеличения добычи нефти и газа на существующих скважинах. Одним из способов достижения этой цели является проведение геолого-технических мероприятий (ГТМ) на скважинах, которые потенциально могут иметь больший дебит. Процедуру подбора ГТМ можно разделить на две крупные задачи: выбор скважин-кандидатов на проведение ГТМ и подбор подходящего ГТМ для каждой выбранной скважины. На большинстве российских и зарубежных нефтегазодобывающих предприятий описанные задачи решаются вручную или с частичной автоматизацией некоторых расчетов, поэтому актуальным остается развитие методов и алгоритмов для их автоматизированного или автоматического решения. Перспективным видится применение методов интеллектуального анализа данных – метода кластеризации данных.

В докладе рассматриваются особенности применения кластеризации для решения задачи выбора скважин-кандидатов на проведение ГТМ, проводится анализ существующих методов и алгоритмов кластеризации данных.

Постановка задачи кластеризации

Математическая постановка задачи кластеризации применительно к выбору скважин-кандидатов на проведение ГТМ из фонда скважин нефтегазодобывающего предприятия может быть выражена следующим образом: на множестве исследуемых скважин n, характеризующимися m геологическими и промысловыми параметрами, провести их разбиение на k кластеров (групп) таким образом, чтобы наиболее похожие скважины принадлежали одному кластеру [1]. Одной из k найденных групп будет искомое подмножество скважин-кандидатов для проведения ГТМ.

Разбиение множества методов кластеризации можно провести в зависимости от выбранного критерия. Более структурированной видится классификация, изображенная на Рис. 1.

Оценим выделенные методы кластеризации и реализующие их алгоритмы с точки зрения их применимости для решения задачи выбора скважин-кандидатов на проведение ГТМ.

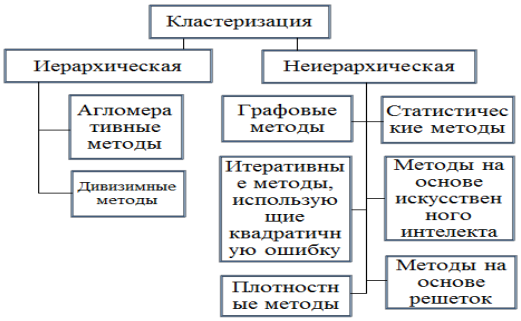

Рисунок 12 - Схема классификации методов кластеризации

Анализ методов и алгоритмов кластеризации

Можно выделить следующие требования, предъявляемые к методам и алгоритмам кластеризации в случае их применения для решения задачи выбора скважин-кандидатов на проведение ГТМ:

1. Нет необходимости предварительного задания числа групп (кластеров), на которое необходимо разбить исходное множество объектов. Это связанно со сложностью определения конкретных групп, а именно их числа и признаков (так, выделение двух групп – скважины-кандидаты и остальные скважины – грубое разбиение, не обеспечивающее приемлемой точности при решении данной задачи).
2. Стабильная работа на одном и том же наборе исходных данных (найденные группы на одном и том же наборе при одинаковых значениях параметров используемого алгоритма должны быть идентичны).
3. Применимость алгоритма на многопараметрических объектах, каковыми являются скважины.

Согласно описанным выше требованиям, применение статистических методов кластеризации (реализуются, например, EM-алгоритмом, который основывается на том, что каждый кластер имеет свое вероятностное распределение, поэтому на исходном множестве объектов образуется смесь распределений, которую необходимо разделить [2]) не подходит для решения задачи выбора скважин-кандидатов на проведение ГТМ, так как требуется задание числа кластеров и для высоко размерных данных сложно сформировать общую функцию распределения.

Анализ также показывает, что итеративные методы (реализуемые алгоритмами K-means, K-medoids, ISODATA) [1], стремящиеся

минимизировать квадратичную ошибку между центрами кластеров и объектами, относящимися к этим кластерам, не удовлетворяют описанным требованиям, так как требуют предварительного задания числа кластеров, а алгоритм K-means не выполняет требование 2.

К числу исключенных алгоритмов можно отнести самоорганизующиеся карты Кохонена, реализующие один из методов на основе искусственного интеллекта, а именно, метода на основе применения нейронных сетей. Данный алгоритм также не удовлетворяет требованию о предварительном задании числа кластеров [3].

Дальнейший анализ с целью выбора подходящих методов и алгоритмов кластерного анализа проведем применительно к иерархическим методам, которые, в свою очередь, подразделяются на агломеративные (восходящие) и дивизимные (нисходящие) [2]. Агломеративные методы характеризуются тем, что в исходном множестве объектов каждый объект считается отдельным кластером. На каждом шаге наиболее схожие объекты и кластеры объединяются, пока все объекты не попадут в один кластер. Для дивизимных методов исходное множество объектов представляется в виде одного кластера, который делится на каждом шаге на составляющие кластеры. В результате работы алгоритмов, реализующих иерархические методы, получается иерархия вложенных объектов, которая может быть удобно представлена в виде дендрограмм – древовидных структур, которые строятся путем пошагового объединения наиболее «похожих» объектов и групп объектов, при этом при объединении «непохожих» по мере сходства (расстояния) объектов (групп объектов) образуется визуальный скачок. В рамках рассматриваемой задачи анализ полученной дендрограммы позволит определить число кластеров, при котором полученные кластеры будут представлять полезную информацию о существующих кластерах скважин на фонде скважин предприятия. Одним из найденных кластеров будет искомый кластер скважин-кандидатов для ГТМ.

Графовые методы кластеризации, которые основываются на теории графов [2], также могут быть применены для решения описанной задачи. Одним из реализующих такие методы является алгоритм MST – минимального остовного дерева, который способен выявлять кластеры произвольной формы и имеет не зависящую от числа параметров временную сложность, что немаловажно для многопараметрических объектов, которыми являются скважины.

Еще одними перспективными видятся плотностные методы кластеризации (например, реализованные в виде алгоритмов DBSCAN, SUBCLU и т.п.), которые позволяют находить сгущения объектов в многомерном пространстве параметров и считать их кластерами [4]. Эти методы находят кластеры произвольной формы и устойчивы к шумам, но, в случае, если искомые кластеры имеют разную плотность, они не будут корректно распознаны. В случае работы с геологическими и промысловыми данными о скважинах и продуктивных пластах, последнее ограничение не существенно, т. к. реальные данные редко образуют кластеры разных плотностей.

И, наконец, рассмотрим применимость методов на основе решеток (реализуются алгоритмами WaveCluster [5], CLIQUE и другими) к решению задачи выбора скважин-кандидатов на проведение ГТМ. Такие методы и алгоритмы могут быть применены к исходному множеству скважин и продуктивных пластов, в силу того, что они не противоречат описанным требованиям, но основной направленностью этих методов является обработка больших объемов исходных объектов, что в случае со скважинами отдельного предприятия может быть излишне.

Заключение

Была рассмотрена возможность повышения уровня автоматизации путем применения методов и алгоритмов кластеризации. Она используется для выявления нетривиальных групп на исследуемом множестве объектов, что может быть использовано для решения задачи выбора скважин-кандидатов на проведение ГТМ. В связи с этим был проведен анализ некоторых методов и алгоритмов кластеризации, что позволило определить из их числа подходящие для решения описанной задачи.

Список использованных источников

1. Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. – 412 с.

2. Айвазян С.А., Бухтштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.

3. Келлер Ю.А. Применение кластеризации данных на основе самоорганизующихся карт Кохонена при подборе скважин-кандидатов для методов увеличения нефтеотдачи // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2014. – №3 (28) – С. 32–37.

4. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, August 2–4, 1996, Portland, Oregon. – The AAAI Press. – P. 226–231.

5. Sheikholeslami G., Chatterjee S., Zhang A. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases // Proceedings of the 24th VLDB Conference, August 24–27, 1998, New York, NY. – Morgan Kaufmann Publishers Inc. – P. 428–439.